Multilingual Word Embeddings for Unravelling the Rule of Law in European Countries

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalezbustamante@fgga.leidenuniv.nl

Jaroslaw Kantorowicz
Leiden University
✉ j.j.kantorowicz@fgga.leidenuniv.nl

Table of Contents

Introduction

This work in progress analyses the salience of different conceptualisations of the ⚖ **rule of law in the parliamentary debates** of 26 European countries from 2015 to 2022 using ParlaMint 4.0 dataset (∼912M words).

We are testing a number of applications of multilingual embeddings (MWEs) to deal with the range of corpora languages (*i.e.*, North and West Germanic, Eastern South and West Slavic, Finnic, Western Romance, among others) and be able to evaluate differences and similarities across countries.

At this stage, we are presenting some bilingual models (BWEs) trained using **machine-translated corpora**.

Different Approaches to the Rule of Law

To what extent should the rule of law exclusively refer to the quality of the legal system (**thin conceptualisation**) and to what extent should it also include human rights and other adjacent concepts (**thick conceptualisation**)?

In a nutshell, this project applies ML to:

- Demonstrate how the meaning of the rule of law has changed over time

- Understanding of the rule of law concept differs across jurisdictions ☑

- Provide novel rule of law indicators

The European Commission for Democracy Through Law (**Venice Commission**) distinguishes the **rule of law**, **Rechtsstaat** and **Etat de droit** as different historical traditions.

The **rule of law** emerges in parallel to the supremacy of the Parliament as a principle of the British unwritten constitution. The rule of law concept, following Dicey's definition, focuses on three core aspects: equality, certainty and non-retrospectively.

However, the concept has evolved including **procedural and substantive aspects**.

Rechtsstaat focused more on the nature of the state and was defined in opposition to absolutism. Therefore, protection had to be provided by the legislative rather than the judiciary.

The French **Etat de droit** focused on the nature of the state as the guarantor of fundamental rights enshrined in the Constitution. Together with **Rechtstaat**, they focused more on the role of the constitution and impulse constitutional review.

Since the end of World War II, there has been a trend to keep some substantive components of the different conceptualisations of the rule of law.

Multilingual Word Embeddings

Parliamentary Debates Data

We use ParlaMint 4.0, which offers data on parliamentary debates of 26 European countries and three Spanish autonomous regions from about 2015 to mid-2022 (some countries even from the 1990s). The dataset offers not only the original speeches but also corpora linguistically annotated in the original languages and machine-translated versions of the texts.
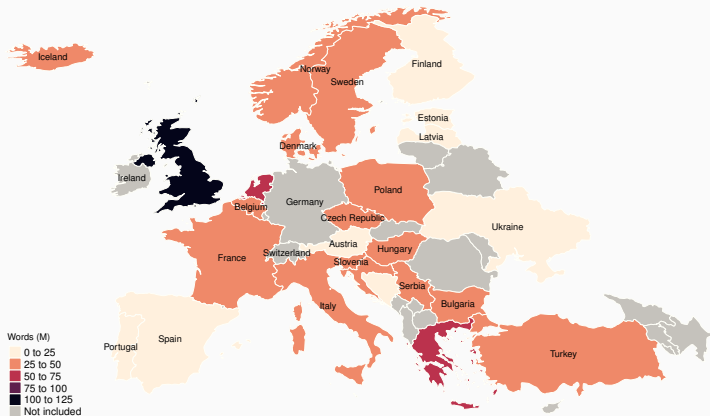
ParlaMint Corpora ~1.1B words
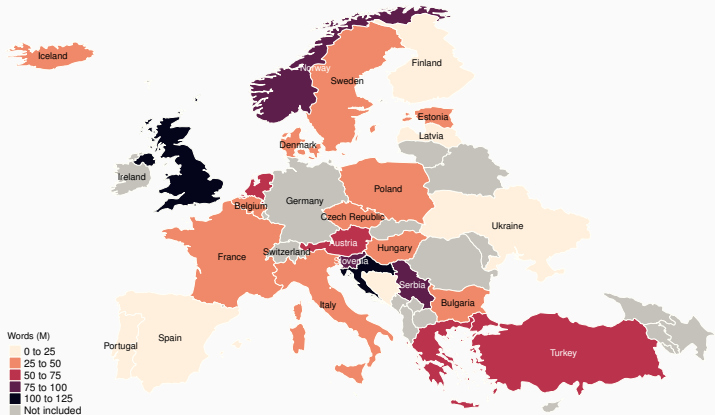
This implies that **~7.5M utterances** were delivered in Parliament in the case of unicameral systems or some of the relevant chambers in bicameral systems.

Austria, Bosnia and Herzegovina, Norway, and Serbia have had coverage since the 1990s. Therefore, to balance better corpora for cross-national comparison, we sliced the data and **focused on the 2015-2022 period**.

Coverage and Machine-Translated Accuracy

🗄 The filtered machine-translated corpora cover ∼**6M** of utterances and ∼**912M words**.

🏛 The upper chamber is not included in some of bicameral systems, namely 🇦🇹 Austria, 🇧🇪 Belgium, 🇨🇿 Czech Republic and 🇫🇷 France. In 🇮🇹 Italy, the *Camera* is not included.

❶ Low machine-translated quality (BLEU ≤ 40) for 🇸🇮 Slovenian, 🇳🇴 Norwegian and 🇧🇦 Bosnian.

📈 In general, the quality of machine-translation is high and fluent ($50 \leq$ BLEU < 60), with several languages standing out with quality that is probably better than an average human's (BLEU ≥ 60). *E.g.*, 🇩🇰 Danish, 🇮🇹 Italian.

* BLEU – Bilingual Evaluation Understudy score

Bilingual Word Embedding Pivot Models

📚 Pre-processing

Tokenization *

Remove punctuation and symbols

Remove numbers and spaces

Use lowercase

Multiword collocations **

* English Punkt tokenizer trained on ~469M

tokens of Wall Street Journal

** Specific list of 40 n-grams: rule_of_law,

open_government, procedural_rights,

justice_access, fundamental_rights,

individual_rights, human_rights, among others.

🖥 Models

Naïve models per country

Procrustes aligned models

</> Parameters

Skim-gram with negative sampling

Training window size of 10

Vectors of 300 dimensions

Five epochs and bootstrapping of 30

Pruning low freq. and noisy words ***

Cython and 12-cores

*** Threshold of 10

Theory-Driven and Experts-Driven Clusters

Procedural		Substantive	
Rules	**Judiciary**	**Democracy**	**Rights**
clarity, stability	judicial_impartiality	democracy	civil_liberties
open_government	judicial_independence	democratic	human_rights
generality, certainty	judicial_protection		fundamental_rights
predictability	judicial_accountability		property_rights
prospective	judicial_review		constitutional_rights
arbitrariness	constitutional_review		minority_rights
publicity	judicial_power		inalienable_rights
transparency	judicial_activism		
congruence, consistency	courts_independence		
coherence	prosecutors_independence		
etc.	etc.		

Preliminary Results

The vocabulary of the naïve models is ~**500K** words. Procrustes alignment helps to deal with **spatial non-comparability**, however, the vocabulary decreases to ~**428K**.
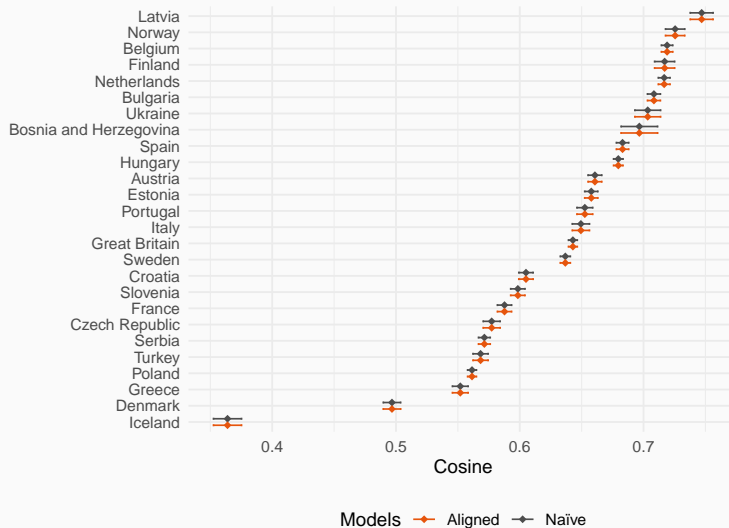
* English as ⚓
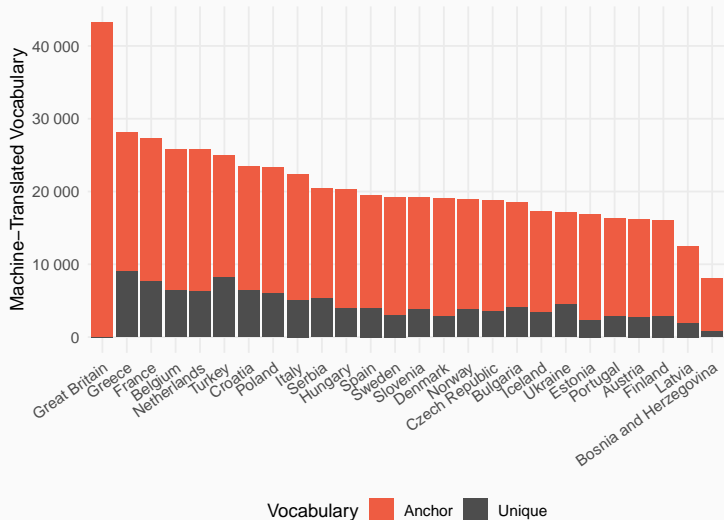
Rule of Law and Procedural Cluster - Judiciary
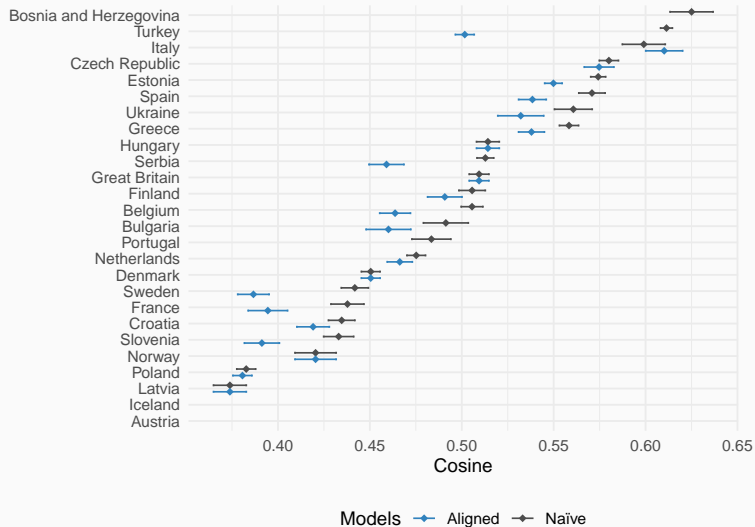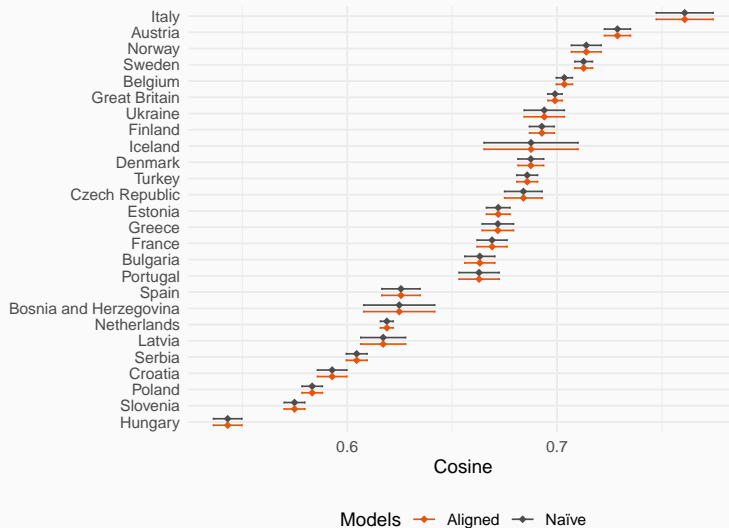
Rule of Law and Procedural Cluster - Rules
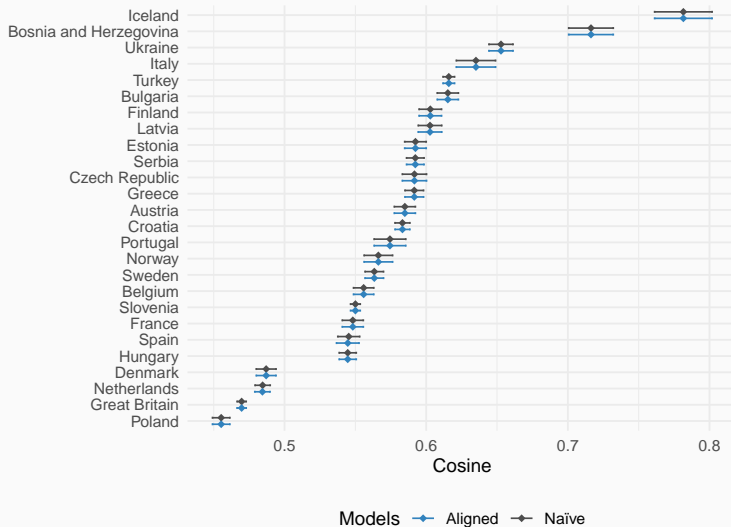
Rule of Law and Substantive Cluster - Democracy

Rule of Law and Substantive Cluster - Rights

Main Patterns of the Rule of Law

Judiciary < Rules < Rights < Democracy ($n = 10$)

🇧🇪 Belgium 🇧🇬 Bulgaria 🇭🇷 Croatia 🇫🇮 Finland 🇱🇻 Latvia 🇳🇱 Netherlands 🇳🇴 Norway
🇸🇮 Slovenia 🇪🇸 Spain 🇺🇦 Ukraine

Judiciary < Rules < Democracy < Rights ($n = 8$)

🇦🇹 Austria 🇩🇰 Denmark 🇪🇪 Estonia 🇫🇷 France 🇮🇹 Italy 🇵🇱 Poland 🇵🇹 Portugal 🇸🇪
Sweden

Judiciary < Democracy < Rules < Rights ($n = 4$)

🇨🇿 Czech Republic 🇬🇷 Greece 🇷🇸 Serbia 🇹🇷 Turkey

Isolated Country-Cases ($n = 4$)

🇧🇦 Bosnia and Herz. (Rules < Democracy < Judiciary < Rights) 🇬🇧 Great Britain (Rules < Judiciary < Democracy < Rights) 🇭🇺 Hungary (Judiciary < Rights < Rules < Democracy) 🇮🇸 Iceland (Judiciary < Democracy < Rights < Rules)

Takeaways and Future Work

Takeaways and Future Work

☑ Most countries tend to exhibit patterns associating the rule of law with more **substantive aspects** such as democracy and rights.

☑ We hope to explore frequency-based approaches to tune our clusters. The judiciary terms seem to have greater **cross-national heterogeneity**, which is noticeable in the alignment and in the differences in this cluster (courts vs. judicial independence).

☑ We want to progress in incorporating texts in the **original languages** into our models and analyses.

❯ In this sense, a potential caveat is that using English as an ⚓ anchor in BWEs could generate underlying bias towards the British definition of the rule of law.

Takeaways and Future Work

☑ We want to explore developing a **tailored gold standard** to evaluate our models and compare their similarities with human coders' opinions on specific concepts.

❯ For the moment, using WordSim-353, the models show average performance except for 🇧🇦 Bosnia and Herzegovina, 🇩🇰 Denmark, and 🇪🇸 Spain (∼0.4). The Bosnian case could be associated with low-quality machine-translation.

☑ We hope to explore the use of **transformers** and domain transfer with LLMs. Also, validate with indicators such as the Rule of Law Dashboard.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com